



AI & The Cyber Security Frontier

SIGS Basel, 17th October 2024

Peter Bury

pbury@paloaltonetworks.com

Topics

- What do you mean, “AI”? ... Themes and Perceptions
- AI Secure Usage
- AI in Cybersecurity

What do you mean, “AI”?

2 Years of GenAI

A Conversation With Bing's Chatbot

Left M

A very stra
search eng

Share full ar



Pausing AI Developments Isn't Enough. We



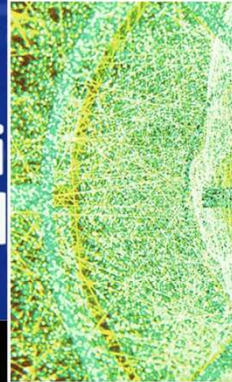
EU Artifi Intell

Microsoft Says Copilot's Alternate

Personality AGI Is an "Ex

"We have implemented

/ Artificial Intelligence / AI Chatbots / AI Applications / Copilot



OpenAI's "Strawberry" Model: Stage 2 Of

GOV.UK

Ali Waseem · Follo
4 min read · Jul 16, 2024

Home > Crime, Justice and Law

2

Press release

UK signs first addressing ris intelligence

The road towards AGI



Lord Chancellor Shab
treaty governing safe

From: [Ministry of Justice](#) and [The](#)
Published 5 September 2024



Palo Alto Networks
1,298,486 followers
2w · Edited ·

We're proud to announce that Palo Alto Networks is among the first 100 companies to sign the EU Artificial Intelligence (AI) Pact!

This initiative moves us closer to shaping AI legislation, driving responsible AI adoption, and safeguarding our digital future.

We're committed to:

- ▶ AI governance: Fostering innovation and preparing for the AI Act.
- ▶ High-risk AI mapping: Ensuring responsible use of AI.
- ▶ Promoting AI literacy: Raising awareness and ethical AI development.

At Palo Alto Networks, we're building a safer, more secure tomorrow.

Let's shape the future of AI together! <https://bit.ly/3Y59b6v>



Emerging Themes, Legislation

AI Factions: Accelerationists vs Altruists

US Executive Order (30/10/23) - Safe, Secure, and Trustworthy Development and Use of AI

The Bletchley Declaration (01/11/23) - An overarching commitment to the design, development, deployment and use of AI in a manner that is safe, human-centric, trustworthy and responsible.

EU AI Act (01/08/2024) Establish a consumer protection-driven approach through a risk-based classification of AI technologies as well as regulating AI more broadly.

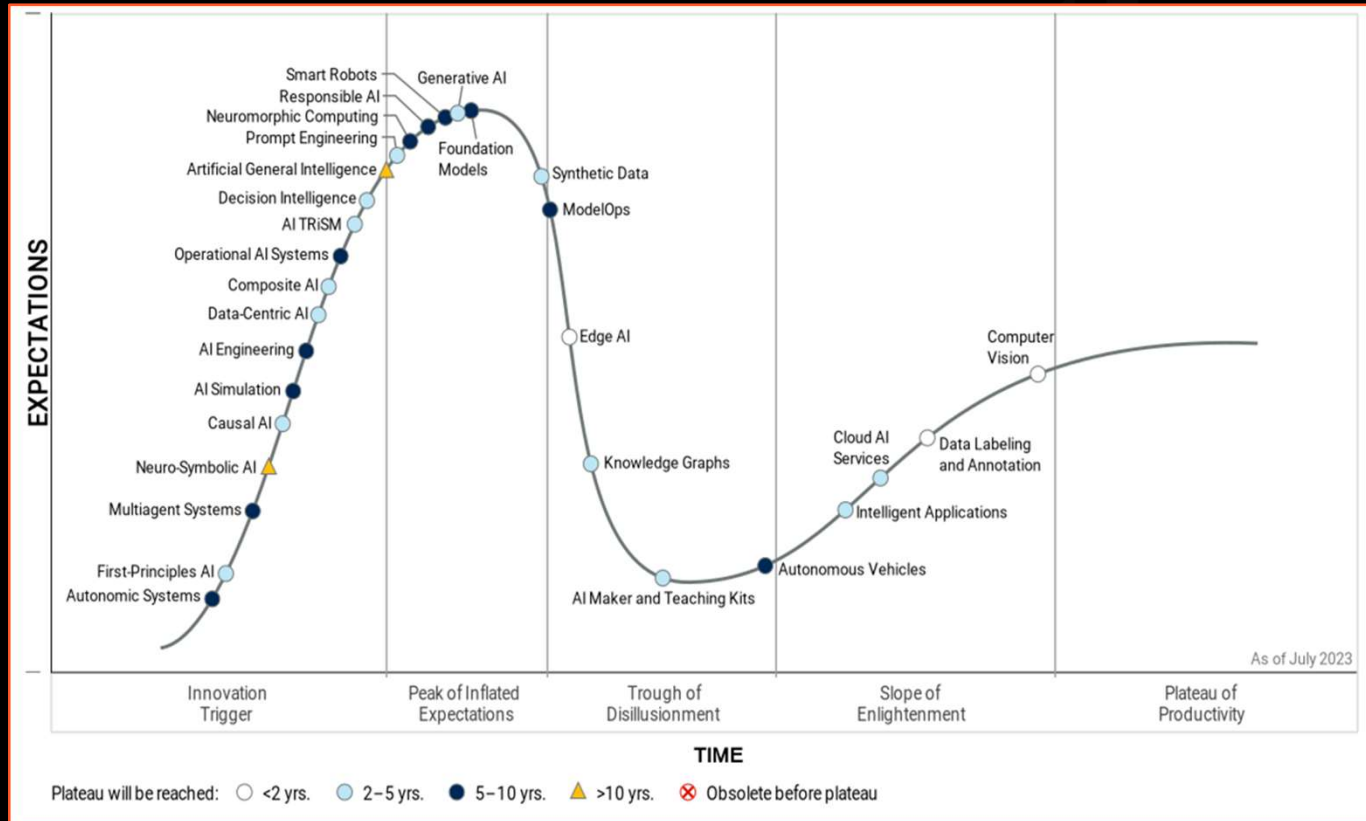
SB-1047 (28/08/2024) Safe & Secure Innovation for Frontier Artificial Intelligence Models Act

U.S. AI Safety Institute/NIST (29/08/24) Agreements enabling formal collaboration on AI safety research, testing and evaluation with Anthropic and OpenAI.

<https://www.aisafetysummit.gov.uk/>

<https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>

Adoption and Integration



Resources - education and talent

The limits of LLMs - Hallucinations?

Transformer Architecture? (Mamba, Megalodon)

Security and Integrity of data in public systems

Legal and copyright

Many organizations lack the data practices necessary to input data into AI

Demis Hassabis, CEO, Google DeepMind

“I would advocate not moving fast and breaking things”

Near Term - Disinformation, bias, fairness, IP and privacy, accelerated cyber threat

AI Proliferation - Bad actors repurposing general purpose technology

AGI - Artificial General Intelligence

@TetraspaceWest

<https://www.lesswrong.com/>

<https://www.theintrinsicperspective.com/p/the-banality-of-chatgpt>

Threats TO AI - Adversarial Attacks

ATLAS™

The ATLAS Matrix below shows the general progression of attack tactics as column headers from left to right, with attack techniques organized below each tactic. & indicates a tactic or technique directly adapted from from ATT&CK. Click on the blue links to learn more about each item, or search and view more details about ATLAS tactics and techniques using the links in the top navigation bar.

Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Collection &	ML Attack Staging	Exfiltration &	Impact &
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection											Cost Harvesting
	Poison Training Data	Phishing &											External Harms
	Establish Accounts &												



- <https://arxiv.org/pdf/2310.13828.pdf>
- <https://arxiv.org/abs/2310.13828>
- <https://atlas.mitre.org/>
- <https://github.com/mitre/advm1threatmatrix>

Threats TO AI - Adversarial Attacks ... And Human Error



Exposed Hugging Face API tokens offered full access to Meta's Llama 2

With more than 1,500 tokens exposed, research highlights importance of securing supply chains in AI and ML

by [Connor Jones](#)

Mon 4 Dec 2023 // 14:00 UTC

The API tokens of tech giants Meta, Microsoft, Google, VMware, and more have been found exposed on Hugging Face, opening them up to potential supply chain attacks.

Researchers at Lasso Security found more than 1,500 exposed API tokens on the open source data science and machine learning platform – which allowed them to gain access to 723 organizations' accounts.

In the vast majority of cases (655), the exposed tokens had write permissions granting the ability to modify files in account repositories. A total of 77 organizations were exposed in this way, including Meta, EleutherAI, and BigScience Workshop - which run the Llama, Pythia, and Bloom projects respectively.

The three companies were contacted by *The Register* for comment but Meta and BigScience Workshop did not respond at the time of publication, although all of them closed the holes shortly after being notified.

The ATLAS Matrix below shows the general categories and links to learn more about each item, or search for specific techniques.

Reconnaissance &	Resource Development &
5 techniques	7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &
Search Victim-Owned Websites	Develop Capabilities &
Search Application Repositories	Acquire Infrastructure
Active Scanning &	Publish Poisoned Datasets
	Poison Training Data
	Establish Accounts &

adapted from from ATT&CK. Click on the blue links to learn more about each item, or search for specific techniques.

ML Attack Staging	Exfiltration &	Impact &
4 techniques	4 techniques	6 techniques
Use Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Use Insecure ML Model	Exfiltration via Cyber Means	Denial of ML Service
Use Insecure ML Model	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Use Insecure ML Model	LLM Data Leakage	Erode ML Model Integrity
		Cost Harvesting
		External Harms



- <https://arxiv.org/pdf/2310.138>
- <https://arxiv.org/abs/2310.138>
- <https://atlas.mitre.org/>
- <https://github.com/mitre/advmlthreatmatrix>

AI and Cybersecurity

Cyber-Enabled vs Cyber-Dependent Crime

~~AI-enabled crimes~~

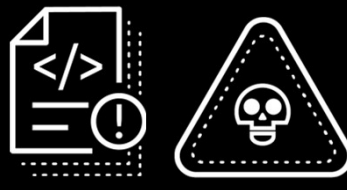
Traditional crimes which can be increased in scale or reach by the use of computers, computer networks or other forms of ICT (such as ~~AI-enabled~~ **AI-enabled** fraud and data theft).

~~AI-dependent crimes~~

Crimes that can be committed only through the use of ~~Artificial Intelligence~~ **Artificial Intelligence** Technology. Where the ~~AI~~ **AI** is (the) ~~device~~ **device**, ~~the device~~ **the device**, ~~and the target of the crime~~ **and the target of the crime**.



Accelerated Attacks



Scaled Attacks



New Vectors

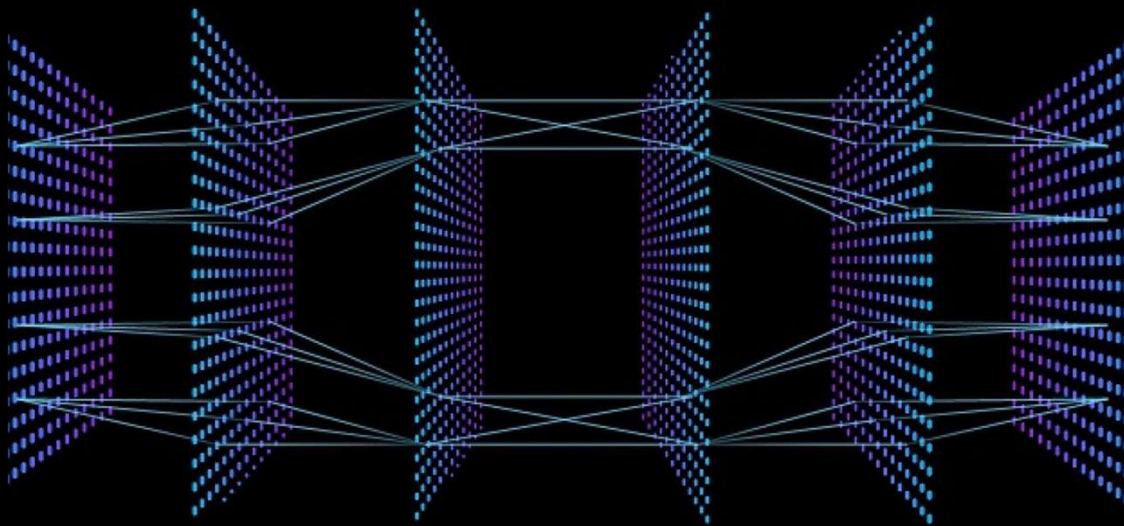
<https://www.cps.gov.uk/legal-guidance/cybercrime-prosecution-guidance>

It's All About Data

Data Gravity: Dave McCrory, 2010

More Data -> Smarter & Better Decisions

More Data -> More Complicated Decisions



<https://datagravitas.com/2010/12/07/data-gravity-in-the-clouds/>

Signatures -> ML -> DL

Machine Learning



- Requires real world threat data to be effective
- Data requires significant curation by data scientists
- Threat features must be identified by humans
- Most effective with a small set of threat features

- A subset of machine learning
- Requires tremendous volumes of real world threat data to be effective
- Data does not require significant curation by data scientists
- Threat features are not identified by humans
- Very effective with large sets of unstructured data

Deep Learning



AI in Cybersecurity - Not all AI is created equal



ANI - Artificial Narrow Intelligence



AGI - Artificial General Intelligence

- Network
- Endpoint



ASI - Artificial Super Intelligence

- Cloud
- Attack Surface



AGI - Artificial General Intelligence



ASI - Artificial Super Intelligence

AI Copilots

- Navigation and Feedback
- Risk Prioritisation
- Best Practice Guidance
- “How-to” answers
- Case Creation + Resolution

• Evolving AI

- *Autonomous Agents*
- *Large Action Models*
- *Objective Driven AI*
- *Alternative Architectures*
- *Retrieval Augmented Generation (RAG)*

AI in Cybersecurity - Not all AI is created equal



ANI - Artificial Narrow Intelligence



- **Precision AI**

- Network
- Endpoint
- Identity
- Cloud
- Attack Surface

- **AI Copilots**

- Navigation and Feedback
- Risk Prioritisation
- Best Practice Guidance
- “How-to” answers
- Case Creation + Resolution

- ***Evolving AI***

- *Autonomous Agents*
- *Large Action Models*
- *Objective Driven AI*
- *Alternative Architectures*
- *Human Copilot?*



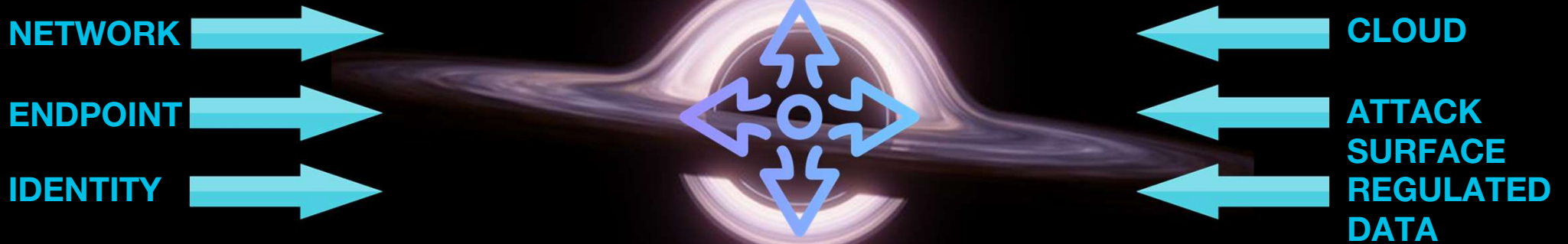
AGI - Artificial General Intelligence



ASI - Artificial Super Intelligence

Data + AI for Better CyberSecurity Outcomes

ML/DL/LLM PRECISION AI



ANALYTICS AND AUTOMATION



GOVERN -> IDENTIFY -> PROTECT -> DETECT -> RESPOND -> RECOVER




AI-ENABLED
THREATS


AI-DEPENDENT
THREATS


AI-TARGETED
THREATS


BUSINESS
THREATS


REGULATORY
RESPONSIBILITIES

<https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf> <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.ipd.pdf>

Operational Resilience, Policy & AI

Top CISO Challenge

CIO DIGITAL INITIATIVES

CISO CYBERSECURITY PROJECTS

CYBERSECURITY TOOLS

Improve Business Intelligence

Data Security

Application Security

Identity & Access Management

Data Loss Prevention, Endpoint Encryption, Data Classification, Cloud Access Security Broker, Single Sign-On, Privileged Access Management

Operational Resilience

Governance, Risk & Compliance

Security Operations

Security Services

Dynamic Application Scanning, Container Security, Static Code Analysis, Web Application Firewall, Identity Management, Multi-Factor Authentication

Process Automation

Cloud Security

Security Operations

Endpoint Security

Risk Monitoring, Supplier / Partner Risk Management, Regulatory / Industry Mandate Compliance, Risk Statistics, Session Replay / Packet Capture

Enhance Customer Presence

Cloud Security

Security Operations

Endpoint Security

Security Monitoring, Digital Forensics, Log Correlation & Analysis, Event Ticketing, User Behaviour Analysis, Malware Analysis

Data Analytics

Cloud Security

Security Operations

Endpoint Security

Treat Intelligence Management, Threat Intelligence, SOAR, Cloud Access Security Broker, System Hardening, System Patching

The typical industry approach requires 10+ point products per digital initiative

Innovate with Emerging Technology

Network Security

Security Services

Endpoint Security

Managed Detection & Response, Endpoint Protection, Cloud Access Security Broker, System Hardening, System Patching

Hybrid-Work Connectivity

Network Security

Cloud Security

Endpoint Security

Endpoint Device Management, Endpoint Encryption, System Hardening, Local Sandboxing, Mobile Threat Detection, Endpoint Protection

Data Quality & Accessibility

Data Security

Application Security

Endpoint Security

Next-Gen Firewalls, Incident Response Services, Attack Surface Management, Threat Research, Managed Threat Hunting, Patching Readiness

Collaborative Workspaces

Content & Collaboration

Cloud Security

Endpoint Security

Secure Web Gateway, DNS Security, Malware Analysis, Encrypted Traffic Management, Email Security, Network Analytics, Intrusion Prevention

Endpoint Security, Cloud Access Security Broker, System Hardening, System Patching, Mobile Threat Protection

Endpoint Device Management, Endpoint Encryption, System Hardening, Local Sandboxing, Mobile Threat Detection, Endpoint Protection

Next-Gen Firewalls, Incident Response Services, Attack Surface Management, Threat Research, Managed Threat Hunting, Patching Readiness

Secure Web Gateway, DNS Security, Malware Analysis, Encrypted Traffic Management, Email Security, Network Analytics, Intrusion Prevention

Endpoint Security, Cloud Access Security Broker, System Hardening, System Patching, Mobile Threat Protection

Endpoint Device Management, Endpoint Encryption, System Hardening, Local Sandboxing, Mobile Threat Detection, Endpoint Protection

Next-Gen Firewalls, Incident Response Services, Attack Surface Management, Threat Research, Managed Threat Hunting, Patching Readiness

Secure Web Gateway, DNS Security, Malware Analysis, Encrypted Traffic Management, Email Security, Network Analytics, Intrusion Prevention

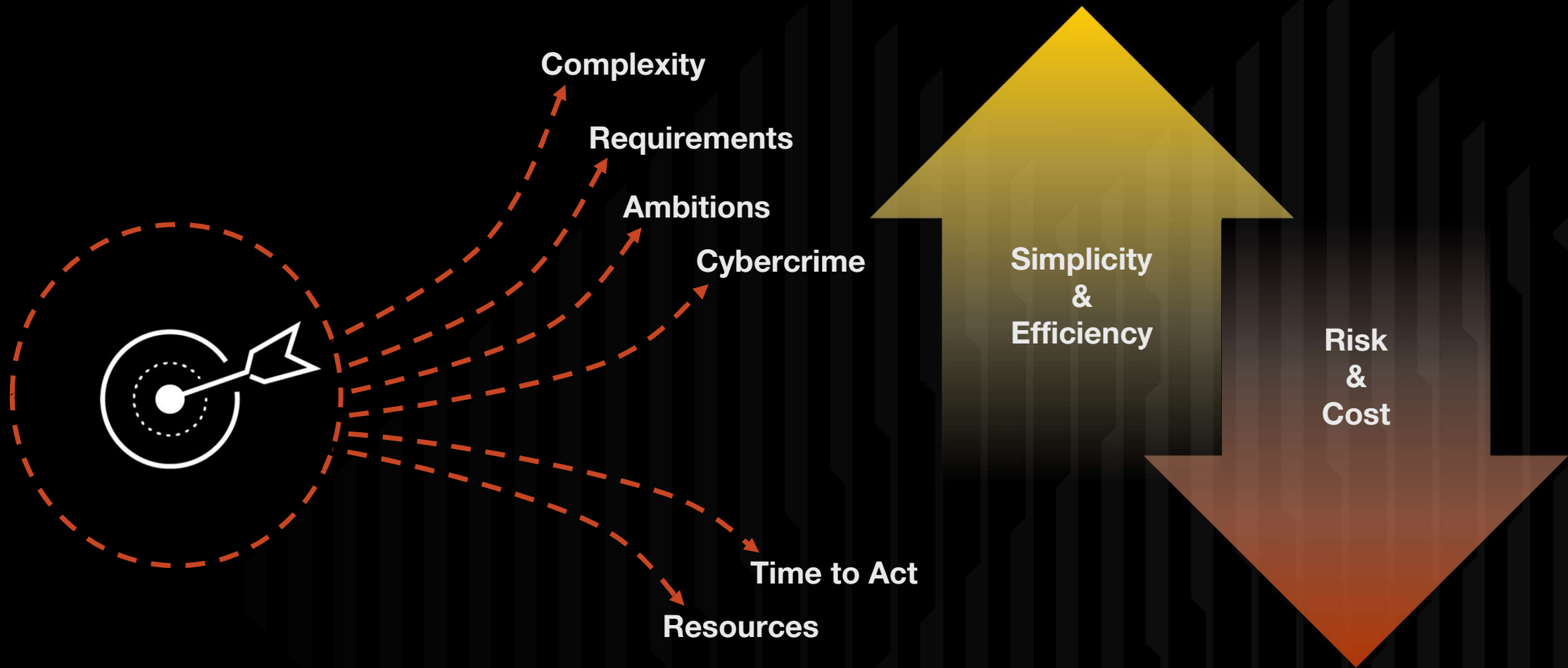
Endpoint Security, Cloud Access Security Broker, System Hardening, System Patching, Mobile Threat Protection

Endpoint Device Management, Endpoint Encryption, System Hardening, Local Sandboxing, Mobile Threat Detection, Endpoint Protection

Next-Gen Firewalls, Incident Response Services, Attack Surface Management, Threat Research, Managed Threat Hunting, Patching Readiness

Secure Web Gateway, DNS Security, Malware Analysis, Encrypted Traffic Management, Email Security, Network Analytics, Intrusion Prevention

The Need for Cyber Resilience



Cyber Resilience Framework



Level 4

- Proactive controls
- Enhanced management
- Highly integrated

Level 3

- Basic controls
- Baseline management
- Some integration

Level 2

- Partially implemented
- Partially managed
- No integration

Level 1

- Not implemented
- Not managed
- No integration

Our Internal Approach to Deploying AI at Palo Alto Networks



**Developing an
AI Policy**



**Determining
AI Exposure
Visibility**



**Protecting
Data Assets**



**Protecting AI
Applications**



**Protecting New HW
and Cloud Scale
for AI**



**Identifying and
Controlling Access
for AI**

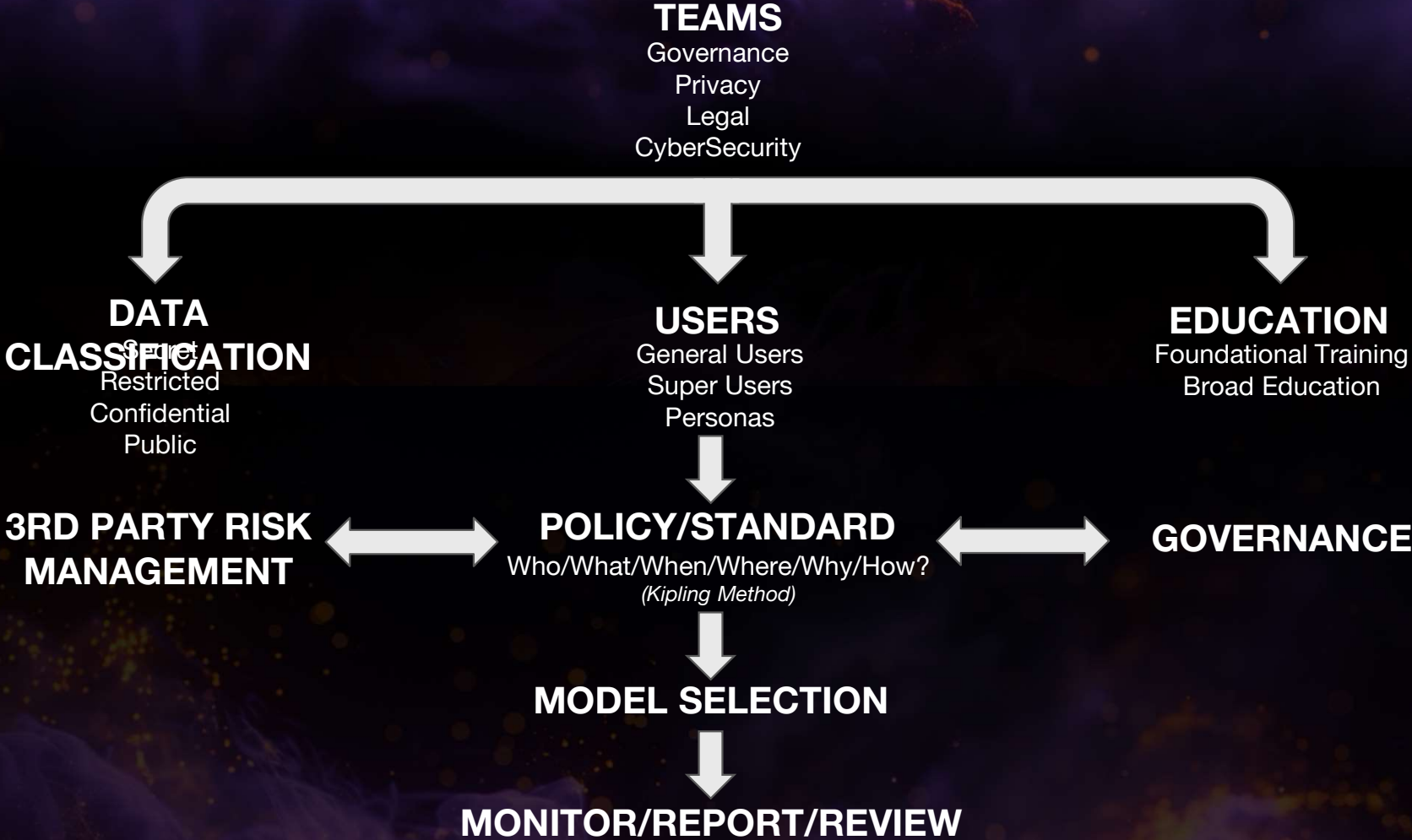


**Protecting
Models**



**Securing the AI
Supply Chain**

Secure AI: Policy & Standards



Secure AI: Visibility and Controls

1 Establish visibility into AI activity
Leverage endpoint data, network data and cloud data

2 Refine Infosec policy
Establish appropriate usage of sanctioned tools/vendors

3 Reduce the exposure
Select a subset of 3rd party LLM SaaS/PaaS providers and bind them with commercial agreements to ensure your data/IP isn't stored or utilized

Repeat & Optimize

5 Implement SOC monitoring
Automate incident remediation

4 Establish controls
Limit and monitor usage to sanctioned tools/vendors

Thank You

CYBERSECURITY
FOR THE AI ERA